Editorial
Special issue on MODEL SELECTION AND HIGH DIMENSIONAL DATA REDUCTION

Nowadays we are often confronted with data sets containing many variables; in some cases the number of variables exceeds the sample size. Indeed, many modern scientific investigations require the analysis of high dimensional data. Examples include genomic and proteomic data, spatial-temporal data, network data, and many others. Modeling such data and in particular high-dimensional data poses many challenges, often involving complex data structures. Specifically, a range of different models with varying complexity can be considered and a model that is optimal in some sense needs to be selected from a set of candidate models. Simultaneous variable selection and parameter estimation play a central role in such investigations. There is now an immense literature on variable selection, and penalized regression methods are becoming increasingly popular. Many new developments have been published in recent years by leading statistical journals.

The application of regression models for high-dimensional data analysis is a challenging task. Regularization techniques have attracted much attention in the literature. Penalized regression is a technique for mitigating difficulties arising from collinearity and high-dimensionality. This approach necessarily incurs an estimation bias, while reducing the variance of the estimator. A tuning parameter is needed to adjust the effect of the penalization so that a desirable balance between model parsimony and goodness-of-fit can be achieved. Different forms of penalty functions have been studied intensively over the last two decades. Examples include the LASSO and its many variants (such as adaptive LASSO, group LASSO, relaxed LASSO, and so on), the SCAD, the Dantzig selector, and the elastic net, to name just a few. More recently, some of these penalization/regularization techniques have been extended to deal with the estimation of large covariance matrices, and the analysis of complex dependence structures such as networks and graphs.

This special issue focuses on computationally efficient strategies, methodology and applications concerning the analysis of complex, high dimensional data set with a focus on model selection and data reduction. The papers contained in this issue deal with submodel selection and parameter estimation for a host of statistical models. Most of the papers are applied in nature and have a computational statistics or data analytic component. We anticipate that the papers published in this special issue will represent a positive contribution to the development of new ideas in the high-dimensional data analysis, and will provide interesting applications. A brief description of the contents of each of the nine papers in this special issue is provided.

The paper by Dernoncourt, Hanczar and Zucker (2013) discusses a feature selection problem in high dimensional data and provides analysis of several real data sets. Vincent and Hansen (2013) developed an algorithm for solving the sparse group lasso optimization problem with a general convex loss function. Furthermore, convergence of the algorithm was established in a general framework. This framework includes the sparse group lasso penalized negative-log likelihood for the multinomial model, which is of primary interest for multiclass classification problems. Beran (2013) considers Hypercube estimators for the mean vector in a general linear model that include algebraic equivalents to penalized least squares estimators with quadratic penalties and to

submodel least squares estimators. He demonstrates that for equal numbers of observations on an array of means , adaptive hypercube estimators simplify greatly: they are then almost identical, for all but small p, to the multiple-shrinkage estimators of Stein (1966). Adaptive hypercube estimators thereby solve a longstanding problem: how to extend Stein (1966) multiple shrinkage to unbalanced designs. Martins and Gabriel (2013) consider model averaging estimation methods in the linear instrumental variables (IV) regression context. The model averaging estimator is a weighted average of individual estimators obtained using different lists of valid instruments. They propose obtaining empirical weights based on existing and well-established instrument selection criteria for IV models and derive some asymptotic properties of the model averaging estimator. Blommaert, Hens and Beutels (2013) propose penalized generalized estimating equations with Elastic Net or L2-Smoothly Clipped Absolute Deviation penalization to simultaneously select the most important variables and estimate their effects for longitudinal Gaussian data when multicollinearity is present. The method is illustrated by mining for the main determinants of life expectancy in Europe. Schomaker and Heumann (2013) propose a framework for model selection and model averaging in the context of missing data. Their focus is on multiple imputation to deal with the missingness and the use of model averaging to incorporate both the uncertainty associated with the model selection and with the imputation process. Monte Carlo simulations reveal the nature of the proposed estimation strategy in the context of the linear regression model. Hall and Xue (2013) investigate a simple recursive approach which, without requiring many extra computational resources, also allows identification of interactions. The suggested strategy can lead to substantial improvements in the performance of classifiers, and can provide insight into how features work together in a given population. Mielniczuk and Teisseyre (2013) propose a random subset method with a new weighting scheme which leads to a novel linear model selection method. They examine its performance on several real datasets. Mougeot and Tribouley (2013) introduce a procedure called Learning Out of Leaders (LOL) and extend it further to adaptively select threshold values (LOLA). The practical performance of LOLA is explored by simulations. Furthermore, using the LOLA algorithm, a solution for modeling the link between the growth rate and the initial level of the gross domestic product is given and the convergence hypothesis is empirically supported.

These nine papers were selected from the many submissions that we received for this special issue. All submitted papers were refereed according to standard procedures for Computational Statistics & Data Analysis. The Guest Editors would like to thank all the authors who submitted their papers for possible publication in this special issue as well as all the reviewers for their valuable input and constructive comments on all submitted manuscripts.

The special issue editors:

S. Ejaz Ahmed
University of Windsor, Canada
E-mail: seahmed@uwindsor.ca

Gerda Claeskens
K.U.Leuven, Belgium
E-mail: Gerda.Claeskens@kuleuven.be

Hidetoshi Shimodaira
Tokyo Institute of Technology, Japan
E-mail: shimo@is.titech.ac.jp

Stefan Van Aelst
KU Leuven, Belgium
E-mail: Stefan.VanAelst@kuleuven.be